CLAIMS

We claim:

1. A method for clustering a set S of n data points to find k final centers, comprising:

partitioning said set S into P disjoint pieces $S_1,...,S_P$;

for each said piece S_i, determining a set D_i of k intermediate centers;

assigning each data point in each piece S_i to the nearest one of said k intermediate centers;

weighting each of said k intermediate centers in each set D_i by the number of points in the corresponding piece S_i assigned to that center; and

clustering said weighted intermediate centers together to find said k final centers, said clustering performed using a specific error metric and a clustering method A.

2. A method according to claim 1 further comprising:

merging said weighted centers into a single dataset D' prior to clustering.

3. A method according to claim 1 wherein P is sufficiently large enough such that each piece S_i obeys the constraint $|S_i| < M$, where M is the size of a physical memory or a portion thereof to be used in processing said each piece.

4. A method according to claim 1 wherein if P is not sufficiently large enough such that each piece S_i obeys the constraint $|S_i| < M$, where M is the size of a physical memory or a portion thereof to be used in processing said each piece, then iteratively performing partitioning, determining, assigning, and weighting until the sets D' of weighted intermediate centers generated thereby obeys the constraint $|D'| < M$.

5. A method according to claim 4 wherein said clustering is performed upon iteratively obtained weighted intermediate clusters.

6. A method according to claim 4 wherein said set S is replaced by weighted intermediate centers of the previous iteration when iteratively performing said partitioning, determining, assigning, and weighting.

7. A method according to claim 1 wherein said determining is performed using said specific error metric and said clustering method A.

8. A method according to claim 1 wherein said specific error metric is the minimizing of the sum of the squares of the distances between points and their nearest centers.

9. A method according to claim 1 wherein said specific error metric is the minimizing of the sum of the distances between points and their nearest centers.

10. A method according to claim 1 wherein said clustering method is an approximation-based method.

11. A method according to claim 8 wherein the distance is the Euclidean distance.

12. A method according to claim 9 wherein the distance is the Euclidean distance.

13. A method according to claim 1 further comprising:

considering a second set of data points for obtaining a second k final centers after said set S is clustered;

repeating partitioning, determining, assigning and weighting for said second set of data points; and

clustering weighted intermediate centers obtained from said second set of data points together with said weighted intermediate centers obtained from said data set S, said clustering performed using said specific error metric and said clustering method A.

14. A method according to claim 1 wherein said partitioning, determining, assigning and weighting is performed in parallel for each piece S_i.

15. An article comprising a computer readable medium having instructions stored thereon which when executed causes clustering a set S of n data points to find k final centers, said clustering implemented by:

partitioning said set S into P disjoint pieces $S_1,...,S_P$;

for each said piece S_i, determining a set D_i of k intermediate centers;

assigning each data point in each piece S_i to the nearest one of said k intermediate centers;

weighting each of said k intermediate centers in each set D_i by the number of points in the corresponding piece S_i assigned to that center; and

clustering said weighted intermediate centers together to find said k final centers, said clustering performed using a specific error metric and a clustering method A.

16. An article according to claim 15 further implemented by:

merging said weighted centers into a single dataset D' prior to clustering.

17. An article according to claim 15 wherein P is sufficiently large enough such that each piece S_i obeys the constraint $|S_i| < M$, where M is the size of a physical memory or a portion thereof to be used in processing said each piece.

18. An article according to claim 15 wherein if P is not sufficiently large enough such that each piece S_i obeys the constraint $|S_i| < M$, where M is the size of a physical memory or a portion thereof to be used in processing said each piece, then iteratively performing partitioning, determining, assigning, and weighting until the sets D' of weighted intermediate centers generated thereby obeys the constraint $|D'| < M$.

19. An article according to claim 1 further implemented by:

considering a second set of data points for obtaining a second k final centers after said set S is clustered;

repeating partitioning, determining, assigning and weighting for said second set of data points; and

clustering weighted intermediate centers obtained from said second set of data points together with said weighted intermediate centers obtained from said data set S, said clustering performed using said specific error metric and said clustering method A, resulting in said second k final clusters.

20. A method according to claim 1 wherein said partitioning, determining, assigning and weighting is performed in parallel for each piece S_i.

21. An apparatus for clustering a set S of n data points to find k final centers, said apparatus comprising:

a main memory;

a processor coupled to said memory, said processor configured to partition said set S into P disjoint pieces $S_1,...,S_P$ such that each piece S_i fits in main memory, said each piece S_i first stored separately in said main memory and then clustered by said processor performing:

for each said piece S_i, determining a set D_i of k intermediate centers;

assigning each data point in each piece S_i to the nearest one of said k intermediate centers;

weighting each of said k intermediate centers in each set D_i by the number of points in the corresponding piece S_i assigned to that center; and

clustering said weighted intermediate centers together to find said k final centers, said clustering performed using a specific error metric and a clustering method A.

22. An apparatus for clustering a set S of n data points to find k final centers, said apparatus comprising:

a main memory;

a plurality of processors coupled to said main memory, one of said processors configured to partition said set S into P disjoint pieces $S_1,...,S_P$ such that each piece S_i fits in main memory, said each piece S_i first stored separately in said main memory and then clustered by each said processor performing:

for each said piece S_i, determining a set D_i of k intermediate centers;

assigning each data point in each piece S_i to the nearest one of said k intermediate centers; and

weighting each of said k intermediate centers in each set D_i by the number of points in the corresponding piece S_i assigned to that center, further wherein after aid weighting, one of said processors finally clustering said weighted intermediate centers together to find said k final centers, said clustering performed using a specific error metric and a clustering method A.